ELSEVIER

International Journal of Forecasting 20 (2004) 69–84

international journal of forecasting

www.elsevier.com/locate/ijforecast

Combining time series models for forecasting

Hui Zou[a], Yuhong Yang[b,*]

[a] Department of Statistics, Sequoia Hall, Stanford University, Stanford, CA 94305-4065, USA
[b] Department of Statistics, Snedecor Hall, Iowa State University, Ames, IA 50011-1210, USA

Abstract

Statistical models (e.g., ARIMA models) have commonly been used in time series data analysis and forecasting. Typically, one model is selected based on a selection criterion (e.g., AIC), hypothesis testing, and/or graphical inspection. The selected model is then used to forecast future values. However, model selection is often unstable and may cause an unnecessarily high variability in the final estimation/prediction. In this work, we propose the use of an algorithm, AFTER, to convexly combine the models for a better performance of prediction. The weights are sequentially updated after each additional observation. Simulations and real data examples are used to compare the performance of our approach with model selection methods. The results show an advantage of combining by AFTER over selection in terms of forecasting accuracy at several settings.

Keywords: Combining forecasts; Forecast instability; ARIMA; Modeling; Model selection

1. Introduction

Let $Y_1, Y_2, \ldots$ be a time series. At time n for $n \geq 1$, we are interested in forecasting or predicting the next value Y_{n+1} based on the observed realizations of $Y_1, \ldots, Y_n$. We focus on one-step-ahead point forecasting in this work.

Statistical models have been widely used for time series data analysis and forecasting. For example, the ARIMA modeling approach proposed by Box and Jenkins (1976) has been proven to be effective in many applications relative to ad hoc forecasting procedures. In a practical situation, in applying the statistical modeling approach, however, one faces the important issue of how to choose the 'best' model among a variety of candidates. Generally speaking, the issue is highly non-trivial and has received considerable attention with different approaches being proposed and studied. We briefly discuss some of these approaches that are closely related to our work. Readers are referred to de Gooijer, Abraham, Gould, and Robinson (1985) for a review on this topic.

For the purpose of selecting a model, the approach of using statistical hypothesis testing techniques has several difficulties. Firstly, one faces the challenging issue of multiple testing, and due to the sequential nature of the tests, there is little one can say about the probabilities of errors associated with the whole procedure after conducting a series of tests. Secondly, there is no objective guideline for the choice of the size of each individual test and it is completely unclear how such a choice affects the forecasting accuracy. In addition,

* Corresponding author. Tel.: +1-515-294-2089; fax: +1-515-294-4040.
E-mail addresses: hzou@stanford.edu (H. Zou), yyang@iastate.edu (Y. Yang).

0169-2070/$ - see front matter © 2003 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.
doi:10.1016/S0169-2070(03)00004-9

even when one compares only two models, the model preferred by a test (or even the true model) does not necessarily perform better than the other one in terms of prediction risk. Alternatively, model selection criteria have been proposed based on different considerations, e.g., AIC (Akaike, 1973) by considering a discrepancy measure between the true model and a candidate, and BIC (Schwarz, 1978) by considering approximating the posterior model probabilities in a Bayesian framework. Hannan and Quinn (1979) proposed a related criterion which has a smaller penalty compared to BIC yet still permits a strong consistency property.

One major drawback with model selection is its instability. With a small or moderate number of observations, as expected, models close to each other are usually hard to distinguish and the model selection criterion values are usually quite close to each other. The choice of the model with the smallest criterion value is accordingly unstable for such a case. A slight change of the data may result in the choice of a different model. As a consequence, the forecast based on the selected model may have a high variability. In a simplified density estimation context, Yang (2001c) showed theoretically the advantage of a proper combining over any selection method.

In this work, we propose the use of a method, named AFTER, to combine the forecasts from the individual candidate models. The idea is that, with an appropriate weighting scheme, the combined forecast has a smaller variability so that the forecasting accuracy can be improved relative to the use of a selection criterion. The method was studied theoretically by Yang (2001b) in a general setting and was shown to have the property that the combined procedure automatically achieves the best rate of convergence offered by the candidates individually. It is unclear, however, how well the theoretical results describe real applications. In this paper, we focus on the ARIMA models, and simulations and some real data sets will be used to compare combining and selection in terms of forecasting accuracy. As will be seen, combining tends to reduce the prediction error when there is difficulty in identifying the best model.

Combining forecasts has been studied for the past three decades (see Clemen, 1989, for a comprehensive review of this topic). Various methods have been proposed. The focus has been on the case where the forecasts to be combined are distinct in nature (i.e., based on very different methods). For example, Clement and Hendry (1998, chapter 10) stated that "When forecasts are all based on econometric models, each of which has access to the same information set, then combining the resulting forecasts will rarely be a good idea. It is better to sort out the individual models—to derive a preferred model that contains the useful features of the original models". In response to the criticisms of the idea of combining, Newbold and Granger (1974) wrote "If (the critics) are saying... that combination is not a valid proposition if one of the individual forecasts does not differ significantly from the optimum, we must of course agree". In our view, however, combining (mixing) forecasts from very similar models is also important. For the reason mentioned earlier, combining has great potential to reduce the variability that arises in the forced action of selecting a single model. While the previous combining methods in the literature attempt to improve the individual forecasts, for AFTER, with the aforementioned theoretical property, it targets the performance of the best candidate model.

Instability of model (or procedure) selection has been recognized in statistics and related literature (e.g., Breiman, 1996). When multiple models are considered for estimation and prediction, the term 'model uncertainty' has been used by several authors to capture the difficulty in identifying the correct model (e.g., Chatfield, 1996; Hoeting, Madigan, Raftery, & Volinsky, 1999). From our viewpoint, the term makes most sense when it is interpreted as the uncertainty in finding the 'best' model. Here *best* may be defined in terms of an appropriate loss function (e.g., square error loss in prediction). We take the viewpoint that, in general, the 'true' model may or may not be in the candidate list and even if the true model happens to be included, the task of finding the true model can be very different from that of finding the best model for the purpose of prediction.

We are not against the practice of model selection in general. Identifying the true model (when it makes good sense) is an important task to understand relationships between variables. In linear regression, it is observed that selection may outperform combining methods when one model is very strongly preferred, in which case there is little instability in selection. In the time series context, our observation is that, again,

when model selection is stable, combining does not necessarily lead to any improvement.

We should also point out that the approach of combining we take is related but different from a formal Bayesian consideration. Particularly, no prior distributions will be considered for parameters in the models.

The rest of the paper is organized as follows. In Section 2, we briefly review some model selection criteria and address some basic issues. In Section 3, the combining algorithm AFTER is presented. In Section 4, we give results of several simulations for comparing combining and selection. Examples of real data are used in Section 5 to demonstrate the advantage of combining by AFTER. Conclusions and discussions are given in Section 6.

2. Some preliminaries

2.1. Evaluation of forecasting accuracy

Assume that, for $i \geq 1$, the conditional distribution of Y_i given the previous observations $Y^{i-1} = \{Y_j\}_{j=1}^{i-1}$ has (conditional) mean m_i and variance v_i. That is, $Y_i = m_i + e_i$, where e_i is the random error that represents the conditional uncertainty in Y at time i. Note that $E(e_i \mid Y^{i-1}) = 0$ with probability one for $i \geq 1$. Let $\hat{Y}_i$ be a predicted value of Y_i based on Y^{i-1}. Then the one-step-ahead mean square prediction error is

$$E(Y_i - \hat{Y}_i)^2.$$

Very naturally, it can be used as a performance measure for forecasting Y_i. Note that the conditional one-step-ahead forecasting mean square error can be decomposed into squared conditional bias and conditional variance as follows:

$$E_i(Y_i - \hat{Y}_i)^2 = (m_i - \hat{Y}_i)^2 + v_i^2,$$

where E_i denotes the conditional expectation given Y^{i-1}. The latter part is not in one's control and is always present regardless of which method is used for prediction. Since v_i is the same for all forecasts, it is sensible to remove it in measuring the performance of a forecasting procedure. Accordingly, for forecasting Y_i, we may define the *net* loss by

$$L(m_i, \hat{Y}_i) = (m_i - \hat{Y}_i)^2,$$

and define the corresponding *net* risk by

$$E(m_i - \hat{Y}_i)^2.$$

Let δ be a forecasting procedure that yields forecasts $\hat{Y}_1, \hat{Y}_2, \ldots$ at times 1, 2 and so on. We consider the average *net* mean square error in prediction

$$ANMSEP(\delta, n_0, n) = \frac{1}{n - n_0 + 1} \sum_{i=n_0+1}^{n+1} E(m_i - \hat{Y}_i)^2$$

from forecasting Y_{n_0+1} up to Y_{n+1}. The observations $\{Y_i\}_{i=1}^{n_0}$ are used for initial estimation and for each $i = n_0 + 1, \ldots, n + 1$, one-step-ahead forecasts are obtained with all the available observations. For the particular case of $n_0 = n$, it is simply the net mean square error in prediction at time n, and will be denoted $NMSEP(\delta, n)$. We will focus on ANMSEP for comparing model selection with model combining in simulations.

For the purpose of comparing forecasting procedures based on real data, given a time series of size n, we will consider the (sequential) average square error in prediction

$$ASEP(\delta, n_0, n) = \frac{1}{n - n_0} \sum_{i=n_0+1}^{n} (Y_i - \hat{Y}_i)^2$$

as a performance measure of a forecasting procedure with n_0 being a fraction (but not too small) of n. Clearly, unlike the theoretical quantity ANMSEP, it can be computed based on the data alone.

In the past few years, there has been interesting work on assessing/comparing predictive accuracy among competing forecasts under general loss and/or possibly non-Gaussian errors; see, for example, Diebold and Mariano (1995), Swanson and White (1995), White (2000), Corradi, Swanson, and Olivetti (2001), and references therein. The proposed approaches can also provide useful information for model comparison.

2.2. Some model selection criteria

Let B denote the backward shift operator, i.e. $BY_i = Y_{i-1}$. ARMA models take the form

$$\phi(B)Y_i = \theta(B)e_i, \tag{1}$$

where $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ and $\theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$ are polynomials of finite orders p and q, respectively. If the dth difference of $\{Y_t\}$ is an ARMA process of order p and q, then Y_t is called an ARIMA(p, d, q) process. ARIMA modeling has now become a standard practice in time series analysis. It is a practically important issue to determine the orders p, d, and q.

For a given set of (p, d, q), one can estimate the unknown parameters in the model by, for example, the maximum likelihood method. Several familiar model selection criteria are of the form

$$-\log(\text{maximized likelihood}) + \text{penalty},$$

and the model that minimizes the criterion is selected. AIC (Akaike, 1973) assigns the penalty as $p + q$; BIC (Schwarz, 1978) as $(p + q) \cdot \ln n/2$; and Hannan and Quinn (1979) as $(p + q) \cdot \ln \ln n$. A small sample correction of AIC for the autoregressive case has been studied by, for example, Hurvich and Tsai (1989). The penalty is modified to be $n(n - p)/(2(n - 2p - 2))$, following McQuarrie and Tsai (1998, chapter 3), where the effective sample size is taken to be $n - p$ instead of n. These criteria will be considered in the empirical studies later in this paper.

The theoretical properties of these criteria have been investigated. It is known that BIC and HQ are consistent in the sense that the probability of selecting the true model approaches 1 (if the true model is in the candidate list), but AIC is not (see, e.g., Shibata, 1976; Hannan, 1982). On the other hand, Shibata (1980) showed that AIC is asymptotically efficient in the sense that the selected model performs (under the mean square error) asymptotically as well as the best model when the true model is not in the candidate list. Due to the asymptotic nature, these results provide little guidance for real applications with a small or moderate number of observations.

2.3. Identifying the true model is not necessarily optimal for forecasting

Suppose that two models are being considered for fitting time series data. When the two models are nested, very naturally, one can employ a hypothesis testing technique (e.g., likelihood ratio test) to assess which of them is more likely to be the one that generated the data (assuming that at least one of the models is correct). Since tests of an exact given size are often hard to find or compute for time series data, asymptotic tests might be used instead. For a small or moderate number of observations, the asymptotic approximation may not be accurate enough. Furthermore, a more serious concern with the approach of attempting to identify the true model for forecasting is that even if one correctly identified the true model, the forecast based on the model actually may not perform as well as a wrong model in terms of forecasting accuracy. As is well known in regression, due to the trade-off between bias and variance, the performance of the true model may be worse than a simpler model. Here we give a simple example, which will be revisited to understand the difference between combining and selection.

Example 0. Consider two simple models: for $n \geq 1$,

$$\begin{aligned} &\text{Model 1}: \quad Y_n = e_n, \\ &\text{Model 2}: \quad Y_n = \alpha Y_{n-1} + e_n, \end{aligned}$$

where $\{e_i\}$ are i.i.d. normally distributed with mean zero and unknown variance σ^2. Here we assume that $0 \leq \alpha < 1$. Obviously, model 1 is nested in model 2 with $\alpha = 0$.

Under model 1, the mean square prediction error is clearly minimized when $\hat{Y}_{n+1} = 0$. If α is actually nonzero, then the NMSEP is $E(\hat{Y}_{n+1} - \alpha Y_n)^2 = \alpha^2 E Y_n^2$. For model 2, with α estimated by $\hat{\alpha}_n$, a natural forecast is $\tilde{Y}_{n+1} = \hat{\alpha}_n Y_n$ and then the NMSEP is $E(\tilde{Y}_{n+1} - \alpha Y_n)^2 = E(\hat{\alpha}_n - \alpha)^2 Y_n^2$. For comparing the performance of the two models, we can examine the ratio

$$\frac{\alpha^2 E Y_n^2}{E(\hat{\alpha}_n - \alpha)^2 Y_n^2}.$$

It does not seem to be easy to examine this ratio analytically. Monte Carlo simulations can be used instead.

Fig. 1 gives a graph of the ratio (on log scale) defined above in α, with $n = 20$ and σ^2 fixed to be 1 based on Monte Carlo simulations with 1000 replications for each choice of α. From the graph, clearly, when α is small (less than about 0.3), the wrong model

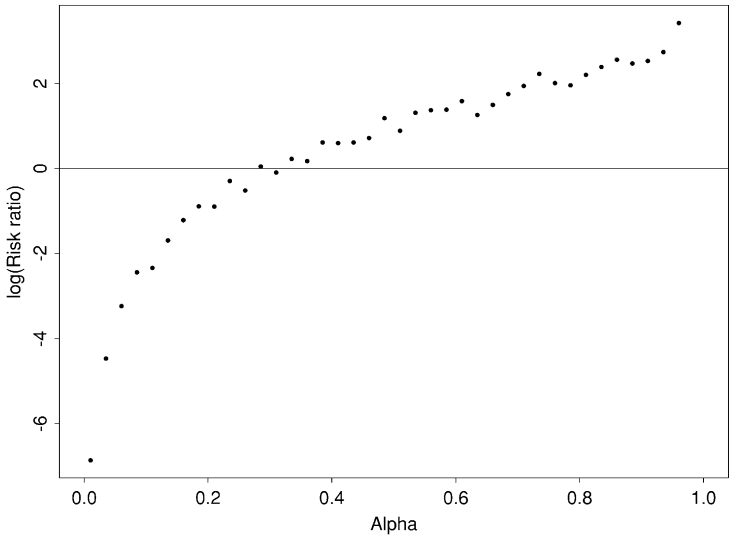

Fig. 1. Comparing true and wrong models in prediction.

performs better. When α becomes larger, however, the true model works better. It is unclear how the significance level of a test relates to which model is better in prediction.

This example was considered by Chatfield (2001, pp. 220–221). He investigated the effect on parameter estimation of α based on a test (in terms of the first-order auto-correlation coefficient) to decide whether α is different from zero. He found through simulations that a substantial bias is introduced in this way.

2.4. Measuring the stability of model selection methods

As mentioned earlier (see also Section 5), model selection can be very unstable. When there is a substantial uncertainty in finding the best model, alternative methods such as model combining should be considered. An interesting issue then is: how to measure the stability of model selection or the resulting prediction? Here we propose two simple approaches.

2.4.1. Sequential stability

Consider a model selection method. One idea of measuring stability in selection is to examine its consistency in selection for different sample sizes. Suppose that the model $\hat{k}_n$ is selected by the method based on all the observations $\{Y_i\}_{i=1}^n$. Let L be an integer between 1 and $n-1$. For each j in $\{n-L, n-L+1, \ldots, n-1\}$, apply the model selection method to the data $\{Y_i\}_{i=1}^j$ and let $\hat{k}_j$ denote the selected model. Then let κ be the percentage of times that the same model ($\hat{k}_n$) is selected, i.e.

$$\kappa = \frac{\sum_{j=n-L}^{n-1} I_{\{\hat{k}_j = \hat{k}_n\}}}{L},$$

where $I_{\{\ \}}$ denotes the indicator function.

The rationale behind the consideration of κ is quite clear: removing a few observations should not cause much change for a stable procedure. The integer L should be chosen appropriately. On one hand, one wants to choose L small so that the selection problems for j in $\{n-L, \ldots, n-1\}$ are similar to the real problem (with the full data observed). Otherwise, even for a stable procedure, the best model at a much smaller sample size can be different from the one at the current sample size. On the other hand, one needs to have L not too small so that κ is reasonably stable. If a method is unstable for the data in the sequential stability, its

ability to pick up the best model is in doubt. We give an example to illustrate this point.

Example 1. Consider AR models up to order 8 as candidates. The true model is AR(2) with coefficients (0.278, 0.366). For $n = 50$ and $L = 20$, based on 1000 replications, the average sequential stability of AIC, BIC, HQ and AICc are 0.70, 0.79, 0.75 and 0.72, respectively, indicating some instability in model selection. Simulations show that the best candidate model (which has the smallest NMSEP among the candidates) is AR(2) for all sample sizes between 30 and 50. Thus it seems clear that here the instability measure κ reflects the difficulty of the model selection criteria in choosing the right model. In contrast, if we consider the same setting as in Example 0 with $\alpha = 0.8$, $n = 50$ and $L = 20$, 1000 replications show that the average sequential stability of AIC, BIC, HQ and AICc are all close to 1 (above 0.99), indicating very little instability in model selection.

It should be pointed out that the measure κ does not address directly the issue of selecting the best model. A model selection method (e.g., the trivial one that always selects the largest model) may well be stable but perform poorly for forecasting.

2.4.2. Perturbation stability

Another approach to measuring stability in model selection is through perturbation. The idea is simple: if a statistical procedure is stable, a minor perturbation of the data should not change the outcome dramatically.

Consider a model selection criterion for comparing the ARMA models in (1). Let $\hat{p}$ and $\hat{q}$ be the orders selected by the criterion. Let $\hat{\phi}(B) = 1 - \hat{\phi}_1 B - \cdots - \hat{\phi}_{\hat{p}} B^{\hat{p}}$ and $\hat{\theta}(B) = 1 + \hat{\theta}_1 B + \cdots + \hat{\theta}_{\hat{q}} B^{\hat{q}}$, where $\hat{\phi}_i$ and $\hat{\theta}_i$ are parameter estimates based on the data. Now we generate a time series following the model

$$\hat{\phi}(B) W_i = \hat{\theta}(B) \eta_i,$$

where η_i are i.i.d. $\sim N(0, \tau^2 \hat{\sigma}^2)$ with $\tau > 0$ and $\hat{\sigma}^2$ being an estimate of σ^2 based on the selected model. Consider now $\tilde{Y}_i = Y_i + W_i$ for $1 \le i \le n$ and apply the model selection criterion to the new data $\{\tilde{Y}_i\}_{i=1}^n$. If the model selection criterion is stable for the data, then when τ is small, the newly selected model is most likely the same as before and the corresponding forecast should not change too much. For each τ, we

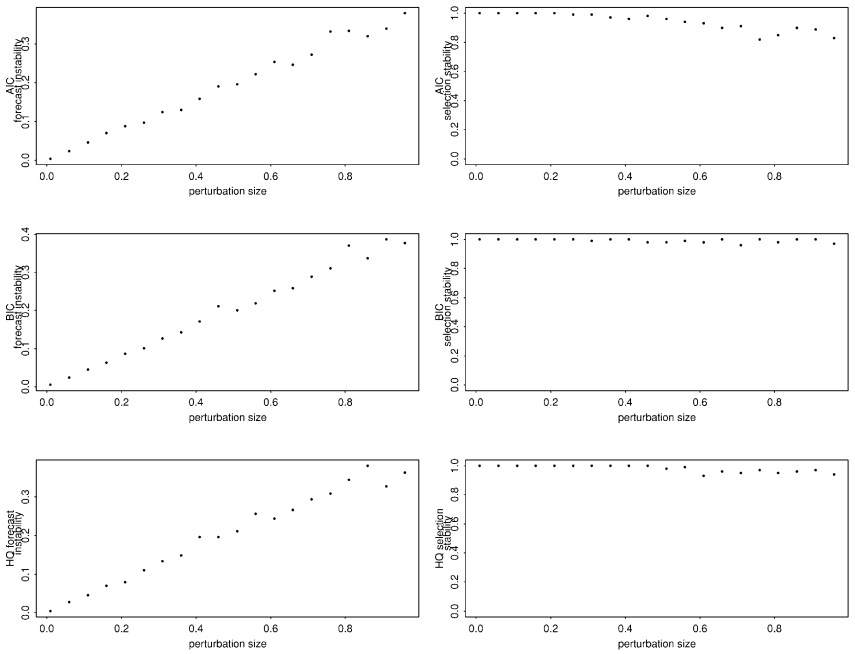

Fig. 2. Forecast and selection stability for data set 1.

generate $\{W_i\}_{i=1}^{n}$ a large number of times (say 100) independently.

2.4.2.1. Stability in selection. For each τ, we record the percentage of times that the originally selected model is chosen again with the perturbation. We then plot the percentage versus τ. If the percentage decreases sharply in τ, it indicates that the selection procedure is unstable even with a small perturbation.

2.4.2.2. Instability in forecasting. Stability in selection does not necessarily capture the stability in forecasting, because different models may perform equally well in prediction. Alternatively, at each τ, we compute the average deviation of the new forecast from the original one relative to the estimated σ (using the initially selected model), based on a large number, say 100, of replications. That is, we average

$$\frac{|\tilde{y}_{n+1} - \hat{y}_{n+1}|}{\hat{\sigma}}$$

over 100 independent perturbations at size τ, where $\tilde{y}_{n+1}$ is obtained by applying the selection procedure again on the perturbed data, and $\hat{y}_{n+1}$ and $\hat{\sigma}$ are based on the original data. It will be called the forecast perturbation instability of the procedure at perturbation size τ for the given data. Again, how fast this quantity increases in τ is a reasonable instability measure.

2.4.3. Data examples

Consider the data sets 1 and 3 (see Section 5 for details).

For data set 1, AR(1) fits very well. Fig. 2 gives the perturbation stability plots over AR models with order up to 5. There is little uncertainty in model selection (AIC, BIC and HQ all select AR(1) and the sequential stability κ is 1 for the selection methods for L not close to n). From the graph, clearly, a small perturbation does not really change the outcome of selection and changes the forecast very little.

For data set 3, a log transformation is used. ARIMA(p,d,q) models with $p = 0,\ldots,5$, $d = 0,1$, $q = 0,\ldots,5$ are the candidates. Fig. 3 shows the

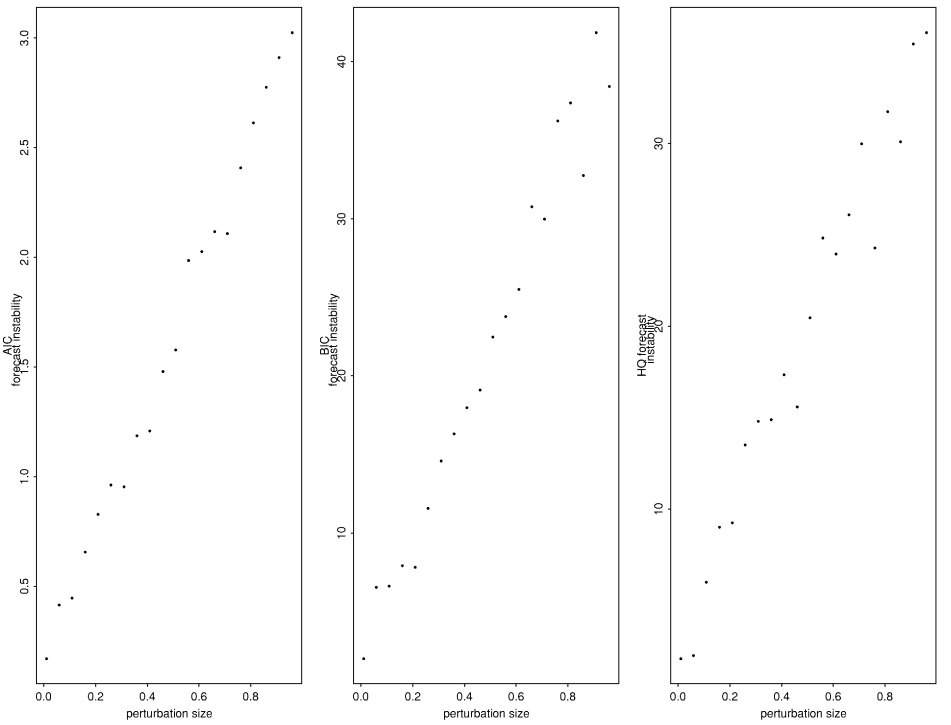

Fig. 3. Forecast instability for data set 3.

forecast instability plots of the model selection methods. For each of AIC, BIC and HQ, the instability in selection increases very sharply in τ.

Compared to data set 1, the perturbation forecast instability of data set 3 is dramatically higher for the model selection methods, especially for BIC and HQ. This suggests that the model selection criteria have difficulty finding the best model. As will be seen later in Section 5, AFTER can improve the forecasting accuracy for this case, while there is no advantage in combining for data set 1. The perturbation stability plots can be useful for deciding whether to combine or select in a practical situation.

In addition to the above approaches, other methods based on resampling (e.g., bootstrapping) are possible to measure model instability. We will not pursue those directions in this work.

3. Algorithm AFTER for combining forecasts

Assume that the conditional distribution of Y_i given $Y^{i-1} = y^{i-1}$ is Gaussian for all $i \geq 1$ with conditional mean m_i and conditional variance v_i. Assume that for each forecasting procedure δ_j, in addition to the forecast $\hat{y}_{j,n}$, an estimate of v_n, say $\hat{v}_{j,n}$, is obtained based on y^{n-1}. If the observations are stationary, various variance estimation methods have been proposed for different scenarios. Note that the procedures do not have to use different variance estimators and if some procedures do not provide variance estimates, we can borrow from others.

Yang (2001b) proposed an algorithm AFTER to combine different forecasts. He examined its theoretical convergence properties. In this work, we apply AFTER to the case that multiple models of the same type are considered for forecasting.

To combine the forecasting procedures $\Delta = \{\delta_1, \delta_2, \ldots, \delta_J\}$, at each time n, the AFTER algorithm looks at their past performances and assigns weights accordingly as follows.

Let $W_{j,1} = 1/J$ and for $n \geq 2$, let

$$W_{j,n} = \frac{\prod_{i=1}^{n-1} \hat{v}_{j,i}^{-1/2} \exp\{-\frac{1}{2}\sum_{i=1}^{n-1}[(Y_i - \hat{y}_{j,i})^2/\hat{v}_{j,i}]\}}{\sum_{j'=1}^{J} \prod_{i=1}^{n-1} \hat{v}_{j',i}^{-1/2} \exp\{-\frac{1}{2}\sum_{i=1}^{n-1}[(Y_i - \hat{y}_{j',i})^2/\hat{v}_{j',i}]\}}. \quad (2)$$

Then combine the forecasts by

$$\hat{y}_n^* = \sum_{j=1}^{J} W_{j,n}\hat{y}_{j,n}.$$

Note that $\sum_{j=1}^{J} W_{j,n} = 1$ for $n \geq 1$ (thus the combined forecast is a convex combination of the original ones), and $W_{j,n}$ depends only on the past forecasts and the past realizations of Y.

Note also that

$$W_{j,n} = \frac{W_{j,n-1}\hat{v}_{j,n-1}^{-1/2} \exp\{-[(y_{n-1} - \hat{y}_{j,n-1})^2/2\hat{v}_{j,n-1}]\}}{\sum_{j'=1}^{J} W_{j',n-1}\hat{v}_{j',n-1}^{-1/2} \exp\{-[(y_{n-1} - \hat{y}_{j',n-1})^2/2\hat{v}_{j',n-1}]\}}. \quad (3)$$

Thus after each additional observation, the weights on the candidate forecasts are updated. The algorithm is called Aggregated Forecast Through Exponential Reweighting (AFTER).

Remarks.

1. Under some regularity conditions, Yang (2001b) shows that the scaled net risk of the combined procedure satisfies

$$\frac{1}{n}\sum_{i=1}^{n} E\left(\frac{(m_i - \hat{y}_i^*)^2}{v_i}\right) \leq c \inf_{j \geq 1}\left(\frac{\log J}{n} + \frac{1}{n}\sum_{i=1}^{n} E\left(\frac{(m_i - \hat{y}_{j,i})^2}{v_i}\right) + \frac{1}{n}\sum_{i=1}^{n} E\left(\frac{(\hat{v}_{j,i} - v_i)^2}{v_i^2}\right)\right),$$

   where c is an explicitly given constant. Thus the risk of the combined forecasting procedure is automatically within a multiple of the best one plus the risk of variance estimation. Consequently, with an appropriate variance estimation, AFTER automatically yields the best rate of convergence provided by the individual forecasting procedures. It is worth noting that the result does not require any of the forecasting procedures to be based on the 'true' model. Similar results in the context of regression can be found in, for example, Yang (2001a).
2. The weighting in (3) has a Bayesian interpretation. If we view $W_{j,n-1}, j \geq 1$, as the prior probabilities on the procedures before observing Y_{n-1}, then $W_{j,n}$ is

the posterior probability of δ_j after Y_{n-1} is seen. However, our procedure is not a formal Bayesian one. No prior probability distributions are considered for the parameters.

3. The normality assumption is not essential. As long as the distribution of the error is known up to a scale parameter, similar combining methods are given in Yang (2001a,b).

4. Simulations

4.1. Two simple models

This subsection continues from Section 2.3. We intend to study differences between selection and combining in that simple setting.

The following is a result from a simulation study with $n = 20$ based on 1000 replications. For applying the AFTER method, we start with the first 12 observations with equal initial weights on the two models. Table 1 gives percentages of selecting the true model by AIC, BIC, HQ and AICc (which are given in Section 2.2) at three representative α values. Table 2 presents the ANMSEPs (as defined in Section 2.1) of the true model, the wrong model, AIC, BIC, HQ, AICc, and AFTER. The numbers in parentheses are the corresponding standard errors of the simulation (i.e., the standard deviation divided by $\sqrt{1000}$).

Figs. 4 and 5 compare the model selection criteria with AFTER in terms of $NMSEP(\delta, 20)$ and $ANMSEP(\delta, 12, 20)$, respectively, based on 1000 replications. Based on Fig. 4, when α is not very small but smaller than 0.55 or so, AFTER performs better than the model selection criteria. It seems natural to expect that, when α is small or big, the selection criteria choose the best model (not necessarily the true model) with very high probability, and this is indeed seen in Table 1. Somewhat surprisingly, AFTER regains an advantage when α is very close to 1 (also shown in Fig. 5). Our explanation is that, for such a case, since the wrong model performs so poorly (over 10 times worse than the true model, as seen in Table 2), even though the probability of selecting the wrong model is very small for model selection (about 10% for $\alpha = 0.91$), the overall risk is substantially damaged. In contrast, by an appropriate weighting, AFTER reduces the influence of the wrong model.

Table 1
Percentage of selecting the true model

	AIC	BIC	HQ	AICc
$\alpha = 0.01$	0.157	0.094	0.135	0.144
$\alpha = 0.50$	0.580	0.463	0.557	0.568
$\alpha = 0.91$	0.931	0.884	0.921	0.924

Table 2
Model selection vs. combining for the two model case in $ANMSEP(\delta, 12, 20)$ and $NMSEP(\delta, 20)$

	True	Wrong	AIC	BIC	HQ	AICc	AFTER
$\alpha = 0.01$							
$n_0 = 12$	0.116	0.000	0.036	0.025	0.036	0.033	0.029
	(0.03)	(0.000)	(0.003)	(0.002)	(0.003)	(0.002)	(0.001)
$n_0 = 20$	0.094	0.000	0.028	0.017	0.025	0.028	0.027
	(0.005)	(0.000)	(0.004)	(0.003)	(0.004)	(0.004)	(0.003)
$\alpha = 0.50$							
$n_0 = 12$	0.158	0.339	0.211	0.234	0.211	0.216	0.158
	(0.005)	(0.007)	(0.005)	(0.005)	(0.005)	(0.005)	(0.004)
$n_0 = 20$	0.124	0.332	0.168	0.192	0.173	0.170	0.156
	(0.008)	(0.015)	(0.011)	(0.012)	(0.011)	(0.011)	(0.010)
$\alpha = 0.91$							
$n_0 = 12$	0.387	5.057	0.944	1.140	0.934	0.988	0.552
	(0.011)	(0.195)	(0.053)	(0.068)	(0.053)	(0.003)	(0.022)
$n_0 = 20$	0.320	5.019	0.614	0.721	0.615	0.615	0.596
	(0.020)	(0.230)	(0.045)	(0.071)	(0.060)	(0.049)	(0.048)

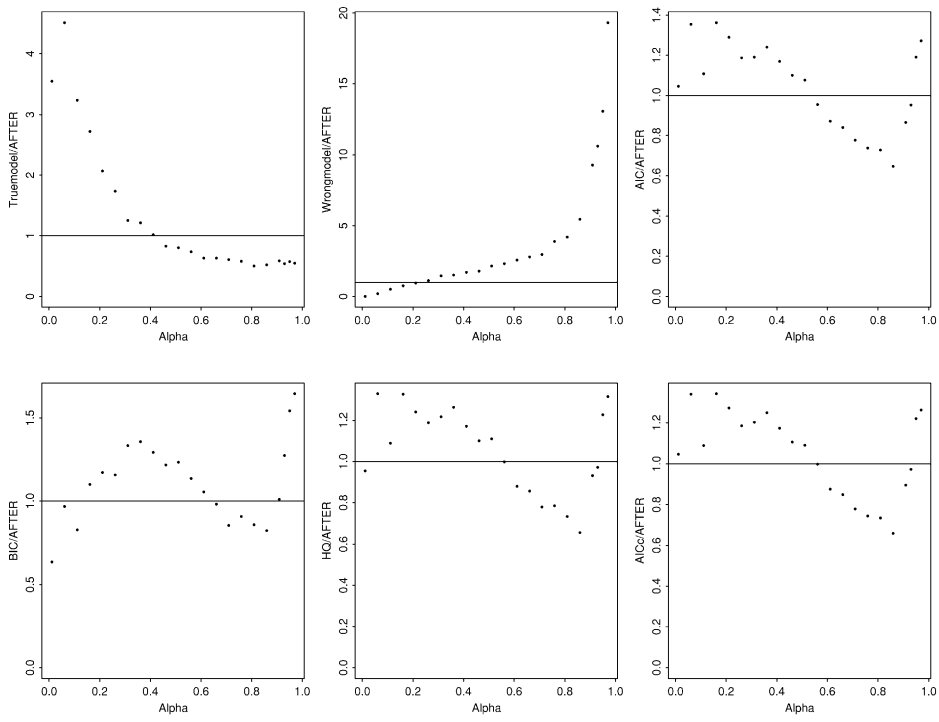

Fig. 4. Comparing model selection with AFTER for the two model case in $NMSEP(\delta, 20)$.

It is interesting to note that, when the average risk $ANMSEP(\delta, 12, 20)$ is considered, the advantage of AFTER is stronger, perhaps because smaller sample sizes are involved in $ANMSEP(\delta, 12, 20)$, where instability in selection tends to be more severe. To confirm this, a simulation was performed on the selection stability at two perturbation sizes: $\tau = 0.2$ and $\tau = 0.6$ (see Section 2.4.2). An additional sample size $n = 50$ was added to show the trend more clearly. Note that, for $\alpha = 0.01$, the selection stability does not change much for n in the range for the model selection criteria. The means of the selection instability for $\alpha = 0.5$ and 0.91 are presented in Table 3 based on 100 replications (data generation) with 1000 perturbations for each data and at each of the two perturbation sizes. The top and bottom entries correspond to $\tau = 0.2$ and $\tau = 0.6$, respectively. The standard deviations are small between 0.02 and 0.04. From Table 3, the selection stability indeed tends to increase in sample size, much more so for $\alpha = 0.91$. Also, not surprisingly, the stability tends to decreases in the perturbation size.

Table 4 gives the average weights that AFTER puts on the two models at the end of 20 observations. It can clearly be seen that, as α increases, AFTER puts higher weight on the right model as is desired.

4.2. AR models with different orders

The candidate models are AR models with orders up to 8. We consider the risk $ANMSEP(\delta, n_0, n)$ with $n = 50$ and two choices of n_0 (20 and 50) for comparing forecasting procedures. Note that the sample size 50 is moderate and we did not find any significant advantage of AICc over AIC (recall that AICc is intended for improving AIC when the sample size is small), and thus AICc will not be considered further.

Five cases are chosen as examples to highlight the comparison between AFTER and model selection criteria. For Case 1, the true model is AR(1) with coefficient 0.8; for Case 2, the true model is AR(2) with coefficients (0.278, 0.366); for Case 3, the true model is AR(3) with coefficients (0.7, −0.3, 0.2); for

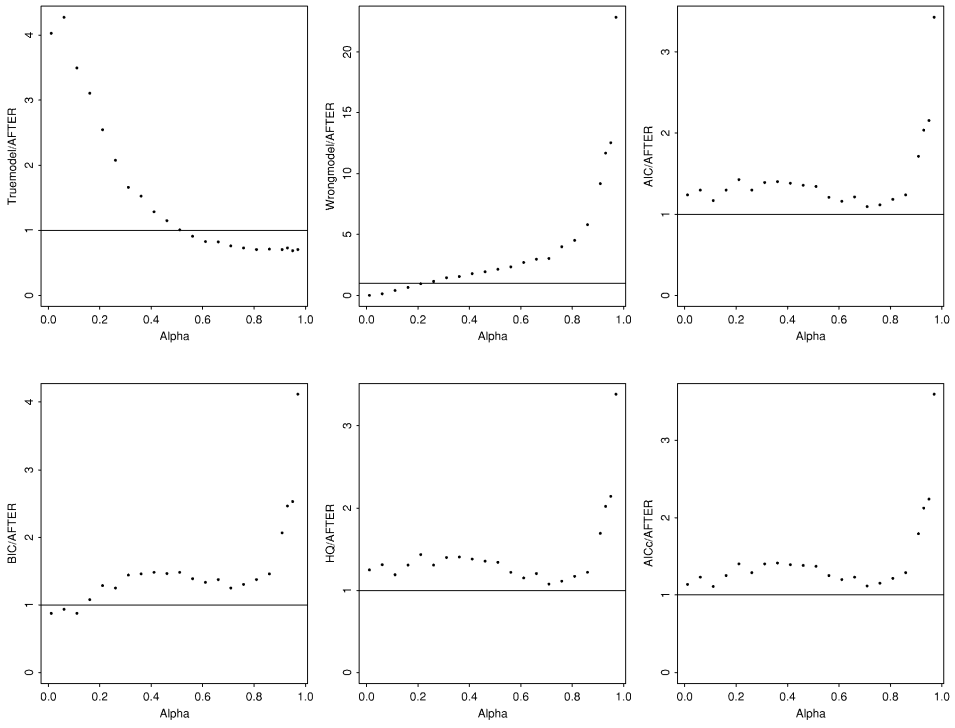

Fig. 5. Comparing model selection with AFTER for the two model case in $ANMSEP(\delta, 12, 20)$.

Case 4, the true model is AR(4) with coefficients (0.386, −0.36, 0.276, −0.227); and for Case 5, the true model is AR(7) with coefficients (−0.957, 0.0917, 0.443, 0.834, 0.323, −0.117, 0.0412). The error variance σ^2 is set to be 1.

Table 3
Selection stability at two perturbation sizes for three sample sizes

	AIC	BIC	HQ	AICc
$\alpha = 0.5$				
$n = 12$	0.867	0.876	0.863	0.871
	0.778	0.824	0.771	0.794
$n = 20$	0.872	0.856	0.865	0.871
	0.802	0.832	0.805	0.809
$n = 50$	0.959	0.920	0.948	0.962
	0.928	0.883	0.912	0.926
$\alpha = 0.91$				
$n = 12$	0.880	0.874	0.874	0.878
	0.775	0.782	0.780	0.790
$n = 20$	0.957	0.926	0.955	0.954
	0.921	0.888	0.918	0.912
$n = 50$	1.000	1.000	1.000	1.000
	1.000	0.999	1.000	1.000

We compare the performances of AIC, BIC, and HQ with AFTER. Two versions of AFTER are considered. The first one starts with the uniform prior weight for forecasting Y_{21} and then updates the weights by the AFTER algorithm in (3). Since the first 20 observations are available in the beginning, when applying AFTER, we can also consider the option of beginning weighting based on the first 15 observations and updating the same way. This way, the use of the uniform prior weighting will have less influence on the forecasts at time 21 and on. They will be referred to as AFTER and AFTER2 in this subsection. In addition, in the competition, the simple averaging method which always assigns equal weight on all the candidate models is also included, denoted

Table 4
Average weights on the models by AFTER

	True	Wrong
$\alpha = 0.01$	0.374	0.626
$\alpha = 0.50$	0.519	0.481
$\alpha = 0.91$	0.793	0.207

by SA. The simulation shown in Table 5 is based on 1000 replications.

For Case 1, the model selection criteria have little difficulty finding the best model and AFTER is not expected to be advantageous. In fact, it performs significantly worse than BIC at both $n_0 = 15$ and $n_0 = 50$. For all the other cases, AFTER performs better than the model selection criteria. The improvements are clearly substantial, with risk reduction up to over one-quarter compared with the best model selection criterion.

Regarding the comparison between the two versions of AFTER, except for Case 1, AFTER2 performs slightly worse than AFTER.

In the simulation, AIC performs better than BIC for some cases, but worse for others. HQ has a risk between those of AIC and BIC, which seems to correspond to the fact that HQ has a penalty (to the likelihood) between AIC and BIC.

Not surprisingly, SA can perform very poorly, with a 50% increase in risk compared to BIC in Case 1. For the other cases, except for Case 4, AFTER performs better than SA. For Case 4, SA has a significantly smaller risk compared to all of the other forecasting procedures, accidentally to a large extent, in our view. Based on the overall performance, it seems clear that AFTER is a better approach for combining forecasts.

4.2.1. Random models

To show that the advantages of AFTER seen above are not atypical, we compare AFTER with model selection in a random setting. We randomly select the true AR order (uniformly between 1 and 8) and then randomly generate the coefficients with uniform distribution on $[-10,10]$ (discard the case if the coefficients do not yield stationarity). One hundred and ten true models were obtained in this

Table 5
Comparing model selection to combining with AR models

	AIC	BIC	HQ	AFTER	AFTER2	SA
Case 1						
$n_0 = 20$	0.164	0.139	0.151	0.161	0.145	0.210
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
$n_0 = 50$	0.107	0.084	0.089	0.092	0.084	0.131
	(0.007)	(0.006)	(0.006)	(0.006)	(0.005)	(0.007)
Case 2						
$n_0 = 20$	0.162	0.174	0.163	0.139	0.139	0.149
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
$n_0 = 50$	0.103	0.122	0.105	0.097	0.097	0.104
	(0.005)	(0.007)	(0.006)	(0.005)	(0.005)	(0.005)
Case 3						
$n_0 = 20$	0.167	0.163	0.164	0.137	0.137	0.149
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
$n_0 = 50$	0.121	0.131	0.125	0.097	0.102	0.098
	(0.006)	(0.006)	(0.006)	(0.005)	(0.005)	(0.005)
Case 4						
$n_0 = 20$	0.194	0.223	0.205	0.152	0.166	0.137
	(0.003)	(0.003)	(0.003)	(0.002)	(0.003)	(0.002)
$n_0 = 50$	0.141	0.184	0.157	0.128	0.141	0.102
	(0.007)	(0.008)	(0.008)	(0.006)	(0.006)	(0.005)
Case 5						
$n_0 = 20$	0.813	0.865	0.835	0.745	0.767	0.833
	(0.012)	(0.013)	(0.013)	(0.012)	(0.012)	(0.014)
$n_0 = 50$	0.604	0.771	0.681	0.593	0.658	0.606
	(0.039)	(0.042)	(0.039)	(0.034)	(0.036)	(0.034)

Table 6
Comparing model selection to combining with AR models: random case

	AIC	BIC	HQ
Median loss ratio	1.347	1.286	1.247
Risk ratio	1.663	1.678	1.578
	(0.092)	(0.133)	(0.099)

way. Table 6 compares AFTER with AIC, BIC, and HQ by examining the ratio of the net risks for $n = 100$ and $n_0 = 50$ based on 100 replications for each model. For AFTER, weighting begins at 50. The losses of the model selection criteria are occasionally much larger compared to AFTER, and the medians of the ratios of the losses of AIC, BIC and HQ compared to AFTER are also given in the table.

The table clearly shows that AFTER performs significantly better in prediction. Figs. 6 and 7 show box plots of the risks of AIC, BIC, HQ and AFTER, and the risks of AIC, BIC, and HQ relative to that of AFTER from the 110 random models.

Note that, based on the 110 random models, AFTER has a smaller risk compared to AIC, BIC, and HQ about 98, 89, and 91% of times, respectively.

4.3. ARIMA models with different orders

Here we compare the model selection methods with AFTER based on ARIMA models with $n = 1$

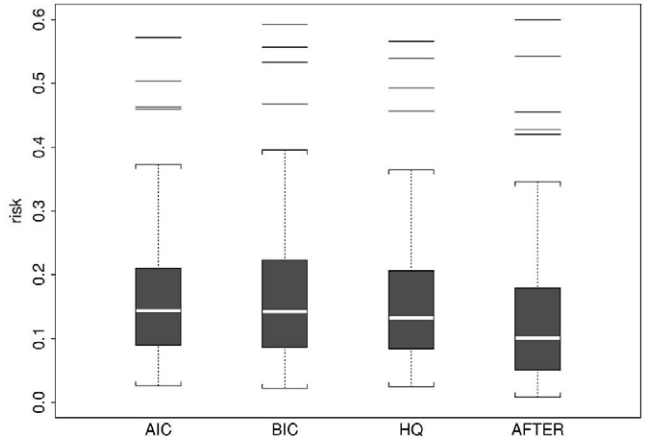

Fig. 6. Risks of AIC, BIC, HQ and AFTER with random AR models.

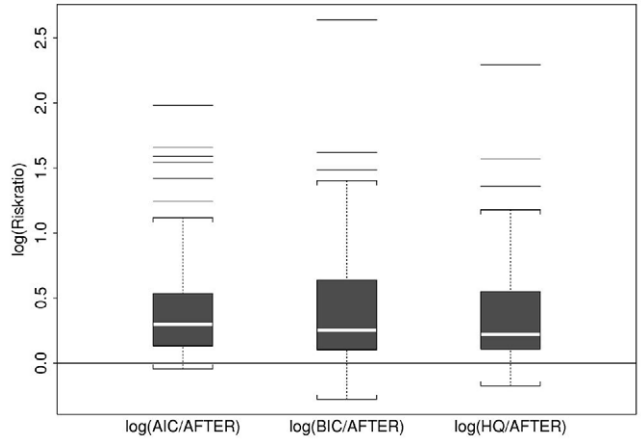

Fig. 7. Risk ratios of AIC, BIC, and HQ compared to AFTER with random AR models.

00 and $n_0 = 50$ with 100 replications. The true model is ARIMA(2,1,2) with AR coefficients (0.8,0.1) and MA coefficients (0.6,0.1). The candidate models are ARIMA(p,d,q) with $p,q = 1,\ldots,8$, $d = 0,1$. Fig. 8 shows box plots of the diffecrences of the net square error losses of AIC, BIC and HQ relative to that of AFTER. Clearly, AFTER performs better on average. In fact, based on the simulation, AFTER has a smaller loss 80, 82, and 78% of times compared to AIC, BIC and HQ, respectively.

5. Data examples

We compared AFTER with model selection on real data. We use $ASEP(\delta, n_0, n)$ as the performance measure. Obviously, n_0 should not be too close to n (so that the ASEP is reasonably stable), but not too small (in which case there are too few observations to build models with reasonable accuracy in the beginning). In this section, AFTER begins with uniform weighting for predicting Y_{n_0+1}.

5.1. Data set 1

The observations were the daily average number of defects per truck at the end of the assembly line in a manufacturing plant for $n = 45$ consecutive business days (see, e.g., Wei, 1990, p. 446). Wei suggests AR(1) model for the data and, indeed, it fits very well. Take $n_0 = 30$ and consider AR models up to order 5. The $ASEP(n_0, n)$ for AIC, BIC, and HQ are all equal to 0.249 and is 0.254 for AFTER. The sequential stability κ ($L = 15$) is 1 for all the selection methods and they all choose AR(1). This is a case where there is little instability in model selection and AFTER has no advantage.

5.2. Data set 2

This data set consists of levels of Lake Huron in feet for July from 1875 to 1972 (see, e.g., Brockwell and Davis, 1991, p. 328) with $n = 98$. Graphic inspections suggested ARMA(1,1) for the data and the residual plots look very reasonable.

Here we consider candidate models ARIMA(p,d,q), $p = 0,1,2$, $d = 0,1$, $q = 0,1,2$. For $n_0 = 78$, the $ASEP(n_0, n)$ for AIC, BIC, and HQ is 0.721, 0.665, and 0.665, respectively. For AFTER, the $ASEP(n_0, n)$ is 0.628, a reduction of 13, 6 and 6%, respectively, compared to the model selection criteria. The sequential stability κ (with $L = 20$) is 0.39, 1, and 1 for AIC, BIC and HQ, respectively. Note that, for different sample sizes between 78 and 98, BIC and HQ always selected ARMA(1,1), but AIC tended to select ARMA(2,1). BIC and HQ are more stable for these data than AIC.

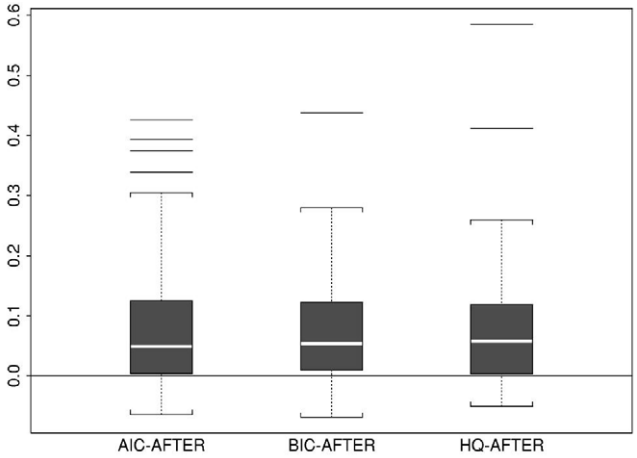

Fig. 8. Loss differences of AIC, BIC, and HQ compared to AFTER for an ARIMA model.

5.3. Data set 3

The data consist of Australian clay brick monthly production statistics (in millions) aggregated into quarters from March 1956 to September 1994 (see, e.g., Makridakis, Wheelwright, & Hyndman, 1998, chapter 1).

Of the $n = 155$ observations, the last 60 were used to assess the performance of the methods. ARIMA(p, d, q) models with $p, q = 0, \ldots, 5$, $d = 0, 1$ are considered as candidate models here. The sequential instability κ (with $L = 60$) for AIC, BIC and HQ is 0.41, 0.26, and 0.17, respectively, suggesting that there is substantial selection instability in the data. The ASEPs for AIC, BIC, HQ, and AFTER are 704.2, 813.4, 785.2, and 635.3, respectively. Note that AIC, BIC and HQ have 11, 28 and 23% higher error compared to AFTER, respectively.

We also considered a log transformation. Although it produced some improvements in diagnostic plots (residuals, ACF, goodness of fit, etc.), model selection is still unstable with κ values for AIC, BIC and HQ of 0.20, 0.40, and 0.36, respectively. The perturbation plot in Section 2.4 also points in this direction. The ASEPs of AIC, BIC and HQ are 14, 29 and 24% higher compared to AFTER for the transformed data.

5.4. Data set 4

The data are monthly sales of new one-family houses in the US from Jan. 1987 through Nov. 1995. (Source: Time Series Data Library, URL: http://www-personal.buseco.monash.edu.au/~hyndman/TSDL.) Of the $n = 107$ observations, the last 20 were used to assess the performance of the methods.

In this data set, an adjustment of the seasonal effects is considered. We decompose the data into seasonal component with period 12 and a remainder series using the `stl` function in Splus. For the remainder series, ARIMA(p, d, q) models with $p, q = 0, \ldots, 5$, $d = 0, 1$ are considered as candidate models. The sum of the forecast of the remainder and the seasonal component gives the final prediction.

The sequential instability κ (with $L = 20$) for AIC, BIC and HQ is 0.11, 0.56, and 0.11, respectively, suggesting that there is a substantial selection instability in the data. The ASEPs for AIC, BIC, HQ, and AFTER are 18.7, 17.4, 16.8, and 12.5, respectively. Note that AIC, BIC and HQ have a 49, 38 and 34% higher error compared to AFTER, respectively.

6. Concluding remarks

Time series models of the same type are often considered for fitting time series data. The task of choosing the most appropriate one for forecasting can be very difficult. In this work, we proposed the use of a combining method, AFTER, to convexly combine the candidate models instead of selecting one of them. The idea is that, when there is much uncertainty in finding the best model as is the case in many applications, combining may reduce the instability of the forecast and therefore improve prediction accuracy. Simulation and real data examples indicate the potential advantage of AFTER over model selection for such cases.

Simple stability (or instability) measures were proposed for model selection. They are intended to give one an idea of whether or not one can trust the selected model and the corresponding forecast when using a model selection procedure. If there is apparent instability, it is perhaps a good idea to consider combining the models as an alternative.

The results of the simulations and the data examples with ARIMA models in this paper are summarized as follows.

1. Model selection can outperform AFTER when there is little difficulty in finding the best model by the model selection criteria.
2. When there is significant uncertainty in model selection, AFTER tends to perform better or much better in forecasting than the information criteria AIC, BIC and HQ.
3. The proposed instability measures seem to be sensible indicators of uncertainty in model selection and thus can provide information useful for assessing forecasts based on model selection.

We should also point out that it is not a good idea to blindly combine all possible models available. Preliminary analysis (e.g., examining ACF) should be performed to obtain a list of reasonable models. Transformation and differencing may also be considered in that process.

Acknowledgements

The authors thank the reviewers for very helpful comments and suggestions, which led to a significant improvement of the paper. This research was supported by the United States National Science Foundation CAREER Award Grant DMS0094323.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N., & Csaki, F. (Eds.), *Proceedings of the 2nd International Symposium on Information Theory*. Budapest: Akademia Kiado (pp. 267–281).

Box, G. E. P., & Jenkins, G. M. (1976). *Time series analysis: forecasting and control* (2nd ed.). San Francisco: Holden-Day.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123–140.

Brockwell, P. J., & Davis, R. A. (1991). Time Series: Theory and Methods. New York: Springer-Verlag.

Chatfield, C. (1996). Model uncertainty and forecast accuracy. *Journal of Forecasting*, *15*, 495–508.

Chatfield, C. (2001). *Time-series forecasting*. New York: Chapman and Hall.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*, 559–583.

Clement, M. P., & Hendry, D. F. (1998). *Forecasting economic times series*. Cambridge University Press.

Corradi, V., Swanson, N. R., & Olivetti, C. (2001). Predictive ability with cointegrated variables. *Journal of Econometrics*, *104*, 315–358.

de Gooijer, J. G., Abraham, B., Gould, A., & Robinson, L. (1985). Methods for determining the order of an autoregressive-moving average process: A survey. *International Statistical Review A*, *53*, 301–329.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, *13*, 253–263.

Hannan, E. J. (1982). Estimating the dimension of a linear system. *Journal of Multivariate Analysis*, *11*, 459–473.

Hannan, E. J., & Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society, Series B*, *41*, 190–195.

Hoeting, J. A., Madigan, D., Raftery, A. E., Volinsky, C. T. (1999). Bayesian model averaging: A tutorial (with discussion). *Statistical Science*, *14*, 382–417.

Hurvich, C. M., & Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika*, *76*, 297–307.

Makridakis, S., Wheelwright, S., & Hyndman, R. J. (1998). *Forecasting: Methods and applications* (3rd ed.). New York: Wiley.

McQuarrie, A. D. R., & Tsai, C. (1998). *Regression and time series model selection*. Singapore: World Scientific.

Newbold, P., & Granger, C. W. J. (1974). Experience with forecasting univariate times series and the combination of forecasts. *Journal of the Royal Statistical Society, Series A*, *137*, 131–165 (with discussion).

Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika*, *63*, 117–126.

Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Annals of Statistics*, *8*, 147–164.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Swanson, N. R., & White, H. (1995). A model-selection approach to assessing the information in the term structure using linear models and artificial neural network. *Journal of Business and Economic Statistics*, *13*, 265–275.

Wei, W. S. (1990). *Time series analysis: Univariate and multivariate methods*. Redwood City, CA: Addison-Wesley.

White, H. (2000). A reality check for data snooping. *Econometrica*, *68*, 1097–1126.

Yang, Y. (2001a). Adaptive regression by mixing. *Journal of American Statistical Association*, *96*, 574–588.

Yang, Y. (2001b). Combining forecasting procedures: Some theoretical results. Accepted by Econometric Theory. Available at http://www.public.iastate.edu/~yyang/papers/index.html.

Yang, Y. (2001c). Regression with multiple candidate models: Selecting or mixing? Accepted by Statistica Sinica. Available at http://www.public.iastate.edu/~yyang/papers/index.html.

Biographies: Hui ZOU received his BS and MS degrees in physics from the University of Science and Technology of China in 1997 and 1999, respectively. He recently graduated from the Department of Statistics at Iowa State University with a MS degree. He is now a Ph.D. student in the Department of Statistics at Stanford University. He has several publications in physics journals, including *International Journal of Modern Physics A*, *Physics Letters A*, *Modern Physics Letters A*, and *Modern Physics Letter B*.

Yuhong YANG received his Ph.D. in Statistics from Yale University in 1996. He then joined the Department of Statistics at Iowa State University as Assistant Professor and became Associate Professor in 2001. His research interests include nonparametric curve estimation, pattern recognition, and combining procedures. He has published several papers in statistics and related journals, including *Annals of Statistics*, *Journal of the American Statistical Association*, *Bernoulli*, *Statistica Sinica*, *Journal of Multivariate Analysis*, and *IEEE Transaction on Information Theory*.